Comparing the costs and acceptability of three fidelity assessment methods for assertive community treatment

Angela L. Rollins (angela.rollins@va.gov)

HSR&D Center for Health Information and Communication, Richard L. Roudebush VAMC

Department of Psychology, Indiana University Purdue University Indianapolis

Marina Kukla (mkukla@iupui.edu)

HSR&D Center for Health Information and Communication, Richard L. Roudebush VAMC

Michelle P. Salyers (mpsalyer@iupui.edu)

Department of Psychology, Indiana University Purdue University Indianapolis

John H. McGrew (jmcgrew@iupui.edu)

Department of Psychology, Indiana University Purdue University Indianapolis

Mindy E. Flanagan (meflanag@iupui.edu)

HSR&D Center for Health Information and Communication, Richard L. Roudebush VAMC

Doug L. Leslie (dleslie@phs.psu.edu)

College of Medicine, Penn State University

Marcia G. Hunt (Marcia.hunt2@va.gov)

VA Northeast Program Evaluation Center, Office of Mental Health Operations

---

This is the author's manuscript of the article published in final edited form as:
Rollins, A. L., Kukla, M., Salyers, M. P., McGrew, J. H., Flanagan, M. E., Leslie, D. L., ... & McGuire, A. B. (2017). Comparing the Costs and Acceptability of Three Fidelity Assessment Methods for Assertive Community Treatment. Administration and Policy in Mental Health and Mental Health Services Research, 44 (5), 810-816. http://dx.doi.org/10.1007/s10488-016-0785-7

Alan B. McGuire (abmcguir@iupui.edu)

HSR&D Center for Health Information and Communication, Richard L. Roudebush VAMC

Department of Psychology, Indiana University Purdue University Indianapolis

Corresponding Author:

Angela L. Rollins, PhD

Associate Director, HSR&D Center for Health Information and Communication

Research Scientist, Richard L. Roudebush VAMC

1481 W. 10th Street (11-H)

Indianapolis, IN 46202

317-988-4789

Author Notes

Acknowledgments: We would like to thank the study participants for making this work possible. Views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs or the U.S. government.

This work was presented at the Twelfth Workshop on Costs and Assessments in Psychiatry, Mental Health Policy and Economics Research: Access, Quality and Outcomes, Venice, Italy (March 27-29, 2015).

Compliance with Ethical Standards

Conflict of Interest: The authors declare that they have no conflict of interest.

Funding disclosures: This work was funded by VA Health Services Research & Development (HSR&D) Grant IIR-09-368. Drs. Rollins, Kukla, and McGuire also were funded by VA HSR&D Grant CIN-13-416. Dr. Kukla also was supported by VA Rehabilitation Research & Development grant D09043. Dr. McGuire was supported by VA Rehabilitation & Development grant D0712-W.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Abstract

Successful implementation of evidence-based practices requires valid, yet practical fidelity monitoring. This study compared the costs and acceptability of three fidelity assessment methods: on-site, phone, and expert-scored self-report. Thirty-two randomly selected VA mental health intensive case management teams completed all fidelity assessments using a standardized scale and provided feedback on each. Personnel and travel costs across the three methods were compared for statistical differences. Both phone and expert-scored self-report methods demonstrated significantly lower costs than on-site assessments, even when excluding travel costs. However , participants preferred on-site assessments. Remote fidelity assessments hold promise in monitoring large scale program fidelity with limited resources.

Key words:
Cost
Fidelity
Quality
Evidence-based practice
Implementation

Comparing the costs and acceptability of three fidelity assessment methods for assertive community treatment

Fidelity to evidence-based models of care has been shown to predict better clinical outcomes for several types of practices (Cuddeback et al., 2013; Drake, Bond, & Essock, 2009; Latimer, 1999; J. H. McGrew, Bond, Dietzen, & Salyers, 1994; McGuire et al., in press; McHugo, Drake, Teague, & Xie, 1999; van Vugt et al., 2011). Fidelity monitoring is a critical tool for providing developmental feedback, either in program start-up for learning program principles and practices or as an important component in supporting the long-term sustainability of such practices (Bond et al., 2014). Assertive community treatment (ACT) is one such evidence-based mental health practice benefitting from fidelity monitoring to develop and maintain quality services according to the model. Targeting consumers who are frequently hospitalized and unlikely to do well in traditional outpatient mental health services, ACT includes a range of intensive, comprehensive services and has been shown to reduce hospitalizations, increase stable housing, and improve engagement in services (Bond, Drake, Mueser, & Latimer, 2001; Bond, McGrew, & Fekete, 1995; Coldwell & Bender, 2007; Herdelin & Scott, 1999; Mueser, Bond, Drake, & Resnick, 1998; Phillips et al., 2001; Ziguras & Stuart, 2000). ACT is best understood as a way of organizing services that fosters integration, teamwork, and continuity of care, and is compatible with other evidence-based practices. ACT uses a multidisciplinary team approach with shared caseloads and daily team meetings to facilitate coordination of care. Team staffing typically includes a psychiatrist, nurse, social worker, employment specialist, substance abuse specialist, case managers, and often a consumer provider. ACT staff provide concrete assistance in the skills of daily living (e.g., budgeting, shopping), primarily in the consumer's home and community. Consumer contact is frequent, with multiple home and community visits per week with each consumer. ACT programs typically serve consumers with the most severe psychiatric disabilities, generally with diagnoses of schizophrenia spectrum or bipolar disorder, concurrent substance use disorders, and extensive hospitalization histories.

Considered the gold standard by many (e.g., (Cuddeback et al., 2013),the Dartmouth Assertive Community Treatment Scale (DACTS) (Teague, Bond, & Drake, 1998) is used extensively for fidelity monitoring by state health authorities. The DACTS includes 28 items assessing team staffing, organizational processes, and services provided by the ACT program. The DACTS has been shown to discriminate between ACT and other services (Teague et al., 1998) and can reliably measure improved model adherence over time (McHugo et al., 2007). However, the DACTS is time intensive, and has traditionally required preparation by the team before the visit and a daylong on-site assessment by an external assessor that includes observations of practice and the daily team meeting, interviews with the team leader and other team staff members and consumers, and a structured chart review. These on-site procedures are too expensive to perform on a large scale, such as in national implementation of ACT or Mental Health Intensive Case Managementwithin the Department of Veterans Affairs (VA).

In an effort to find reliable, valid, and lower cost fidelity assessment methods, two forms of remote fidelity assessment for ACT have been developed and have shown promise: phone-based and expert-scored self-report (J. H. McGrew, White, Stull, & Wright-Berryman, 2013; Rollins et al., 2016). In expert-scored self-report, teams complete structured data tables detailing team composition, functions, and patient care that are then coded and analyzed by expert reviewers who assign a score on corresponding fidelity items. Phone-based fidelity assessments also use the same data table preparation by the team but add a phone call review between fidelity assessors and the team leader to review the prepared data for clarifications and context. In the most recent study using 32 VA MHICM teams, investigators found good inter-rater reliability on the DACTS total mean score for both phone and expert-scored self-report in terms of both consistency between raters, measured by two-way mixed intraclass correlation coefficient (ICC=.96 for each using average measures; ICC=.92 for each using single measures). Also inter-rater consensus was measured by the mean absolute difference between raters: .09 for phone and .10 for expert-scored self-report, which reflect small discrepancies on

a scale with possible scores ranging from 1.00 to 5.00. Both remote methods showed good concurrent validity on total DACTS mean score in terms of both strong inter-method consistency with on-site score (ICC=.91 for phone and ICC=.84 for expert-scored self-report, both using average measures), and inter-method consensus between on-site and remote assessments measured by the absolute mean difference with on-site score (.11 for phone and .17 for expert-scored self-report) (Rollins et al., 2016). The purposes of the current study were to 1) document and compare the costs of each fidelity assessment method, and 2) to assess the acceptability of each method from the perspective of ACT providers.

Methods

Thirty-two VA mental health intensive case management teams participated, each receiving a phone-based, expert-scored self-report, and an on-site fidelity assessment with experienced fidelity assessors using the DACTS. Sites completed a detailed set of preparation tables and then were assigned to receive first either a phone or on-site visit, scheduled within two weeks of one another. We used stratified random sampling based on type of VA facility (20 general medical and surgery and 12 neuropsychiatric facilities) and the team's previous performance (high vs low performing using a median split) on a fidelity self-assessment used in routine practice in VA ("pure" self-report as opposed to expert-scored self-report). The order of phone and on-site assessments were counter-balanced because the first assessment influenced time required for the second assessment. For example, completion of the phone assessment preparation tasks would decrease the amount of preparation time the site would spend on their subsequent on-site assessment. However, 19 sites received on-site assessments first and 13 sites received the phone assessment first, due to 3 sites requiring last minute rescheduling of their phone-based assessment.

Team leaders were the primary study respondents for scheduling assessments, participating in phone assessments, and providing tabled self-study information prior to assessments. However, team leaders were encouraged to use discretion in enlisting other staff

to help in completing the 10 fidelity preparation tables that addressed team structures, activities, and consumer care characteristics, such as team composition, census of team veterans receiving various types of care, and circumstances for recent veteran admissions to the team, discharges from the team, and hospitalizations.The 10 tables were used for all 3 assessments.

Two raters rotated responsibility for on-site or phone-based assessments across the 32 teams. Three additional raters alternated as the second phone-based fidelity rater and as one of the two raters for expert-scored self-report assessments. Expert-scored self-report raters scored the DACTS independently after reviewing tabled information from the team and came together later for consensus ratings. Phone raters received the prepared tables and held a phone call with the team leader to review each completed table and ask for clarifications if needed. One phone rater led the phone assessment interview while the other asked additional questions if needed. All raters were experienced with fidelity assessment prior to the study and had either master's level or doctoral degrees. In addition, all raters underwent an all-day training to review DACTS fidelity assessment protocols adapted for the VA services context and participated in monthly calls to maintain consistency of methods.

Only costs for the method first delivered (phone or on-site) were used in this study. Expert-scored self-report was not impacted by order of administration because it used the site's prepared information for the "first" assessment. Thus all 32 observations were included in time and cost estimates for expert-scored self-report. The first step in our cost estimation calculated the average time devoted by two assessors. In the case of phone and expert-scored self-report, two assessors were required to document inter-rater reliability of those methods, a key aim of an earlier study (Rollins et al., 2016). Two raters were used in 4 cases for the on-site method as a confirmation of reliability. The use of single rater cost data provided a parallel measure of assessor time for all assessment cases and types. Also, use of single rater data reflected real-world use of one assessor for large implementation efforts outside of rigorous research protocols. For the costs of other participants in the fidelity process, assessors recorded the time

and staff person(s) participating in each element of the assessment, and team leaders reported staff names and time spent in pre-assessment table preparation activities (common to all three assessments). VA staff salaries are published on public websites and were recorded in our datasets. To compute personnel costs, we used the following formula: Minutes in fidelity activity * (base salary per minute + fringe rate assumed at 30% of base). Assessors also recorded travel costs for each on-site visit (e.g., airfare, hotel, ground transportation). After computing descriptives, we used independent t-tests to make comparisons between phone (cost data collected in 13 sites) and on-site assessments (collected in the other 19 sites). For all comparisons involving expert-scored self-report assessments, we used paired t-tests to detect cost differences because expert-scored self-report assessments represented repeated observations of a site (phone and self-report paired comparisons for 13 sites; on-site and self-report paired comparisons for 19 sites). T-values and degrees of freedom were adjusted using the Satterwaithe correction when variances between groups were not equivalent.

After teams completed all three fidelity assessments, research assistants interviewed team leaders to gather feedback and preferences regarding the fidelity assessment methods. Interviews were audiotaped, transcribed, and entered into Atlas-ti qualitative data analysis software to facilitate coding and accessing data. The interviews were reviewed using an iterative, consensus-building process to identify emergent themes related to barriers and facilitators of fidelity assessment, consistent with content analysis techniques described by Crabtree and Miller (1999). Initially, a senior researcher and a research assistant independently read interviews to identify barriers, facilitators, and preferences of the fidelity assessment methods. After this initial step, the raters met regularly to discuss findings and develop a working set of codes. This process was repeated on fresh sets of interviews until a set of defined codes was identified. Then, all interviews were coded in blocks of three (two independently and the third in common) to maintain consistency. During this intensive coding process, raters met weekly to compare coding of the common interviews, resolving

discrepancies through consensus. For this analysis, documents were grouped by fidelity assessment type, and we compared preferences expressed for each assessment, and reasons provided for those preferences.

Results

Data regarding personnel time and other costs are presented in Table 1. Costs for the on-site assessments that were administered first (n=19) averaged US$2579, with US$1663 in personnel costs and US$916 in travel costs. Costs for the phone assessments that were administered first (n=13) were US$571 and all expert-scored self-report assessment methods (n=32) averaged US$553. Phone assessment costs were significantly lower than on-site assessment costs ($t(21)=10.4$, $p< .01$). Similarly, using paired t-tests, the expert-scored self-report cost was significantly lower than the onsite assessment cost ($t(18)=13.7$, $p< .01$) and the phone assessment cost ($t(12)=8.8$, $p< .01$). Excluding travel costs which inflate the cost of on-site assessments, total personnel costs for phone were still significantly lower than for on-site assessment costs ($t(24)=7.5$, $p<.001$). Likewise, total personnel cost for expert-scored self-report was significantly lower than for on-site assessment cost at those sites ($t(18)=14.2$, $p< .01$). The responding teams' personnel costs for the on-site assessments ($782) were significantly higher than for phone assessments ($439; $t(25)=2.7$, $p=0.01$). Similarly, using paired t-tests, the responding teams' personnel cost for the on-site assessment was significantly higher than for expert-scored self report ($488; $t(18)=10.2$, $p<.01$). We conducted some sensitivity analyses to account for the wide range of salaries in our assessors. We compared differences in cost by applying lowest salary (Master's degree-level, likely to reflect real-world fidelity monitors) to all assessor time. Resulting cost differences remained significant, even when applying a uniform, modest salary cost.

Qualitative interviews revealed that most respondents (75%) expressed a preference for the on-site assessment methods over remote methods. Reasons for this preference included the following: assessor traits, such as being knowledgeable regarding the clinical model; greater

perceived accuracy for on-site assessments (e.g., easier to communicate about program in person, assessor able to "see" program in action); the personal contact of in-person visits; and getting informal feedback throughout the visit, particularly from an "outsider." For instance, one respondent stated, "[on-site assessment] really engages the teams and makes them perform." Another participant noted, "[on-site assessment] adds weight and seriousness to the [fidelity assessment] process…Filling out a paper doesn't have the same impact as someone coming to visit." Directly comparing fidelity methods, another commented, "[the assessor] could see more of the process first hand, that would have been lost over the phone." Although most favored on-site visits, when phone visits were preferred, it was usually because they required less time. For example, a participant remarked, "Phone assessment cuts down time and because it is briefer… it interferes less with [clinical] programming than the on-site [fidelity assessment]." Indeed, the most frequent disadvantage to on-site visits was the length of time required; for example, one participant noted "It was a busy day and took time away from Veterans, but it was not too intrusive."

Discussion

In this study, phone and expert-scored self-report costs were considerably less than on-site assessment costs: on-site costs were almost five times those of remote fidelity assessment methods. Cost savings realized by the remote fidelity assessment methods appear to be primarily attributable to assessor personnel and travel costs. The cost burden to the teams participating in the fidelity assessment differed statistically across fidelity assessment types, but were within $300-$400 of one another. Although costs to perform phone fidelity assessments were statistically higher than those for expert-scored self-report, the practical significance of these differences is negligible ($18), and may be vulnerable to limited sample size for phone fidelity assessments (n=13). Given that the phone assessment was more strongly associated with the on-site method than was the expert-scored self-report (.96 vs. .84) (Rollins et al., 2016), the return on investment was $15/.1 increase in ICC.

Despite cost savings of remote methods, however, most respondents expressed a preference for the on-site assessment methods. On-site assessments have the advantage of in-person, one-on-one interaction between the assessor and program leadership, which respondents perceived as leading to better understanding of their program and better ability to receive and use real-time feedback. Although remote fidelity assessments showed favorable reliability and had high concurrent validity with the on-site method in previous work (Rollins et al., 2016), the perception of being understood, heightened perhaps in on-site interactions, could be important for recipients of fidelity assessment. For instance, feeling understood may lead staff to be more open to corrective feedback based on the fidelity assessment. This is a clear advantage when the feedback from fidelity assessment is perceived as important (e.g., for program improvement), in addition to the scores themselves. Participants generally viewed these intensive interactions as beneficial to understanding program strengths and areas of deficiency, including clear communication of action steps toward improvement and better adherence to the program model. Moreover, while on-site visits were more time intensive, detracting from clinical duties, the majority of participants viewed this trade-off as acceptable given its many advantages. Furthermore, although feedback during the phone fidelity assessments was not part of our protocol in this study, these findings suggest that real time feedback could be a valuable addition to future implementation of phone fidelity assessments, possibly attenuating some of the preferences expressed for on-site assessments.

Conclusions and Future Directions

Combined with the good inter-rater and concurrent validity found in previous work (Rollins et al., 2016), these cost data support remote fidelity assessments as promising alternatives to on-site methods. However, our qualitative findings also highlight the value of the on-site fidelity assessment for the technical assistance and expert guidance that accompanies a review. In line with this finding, previous research has found that traditional on-site fidelity monitoring serves as a key support to staff implementing evidence based practices in the

community, resulting in lower staff turnover over time (Aarons, Sommerfeld, Hecht, Silovsky, & Chaffin, 2009). Taken together, these findings suggest an important role for on-site fidelity assessment in select contexts. For example, remote fidelity assessment methods could be incorporated into a stepped protocol for fidelity assessment, as suggested elsewhere (John H. McGrew, Stull, Rollins, Salyers, & Hicks, 2011). In a stepped approach, on-site assessments would be reserved for sites with "trigger" events, such as new team formation, significant staff turnover, or turnover in critical positions (e.g., team leader), and teams experiencing low fidelity scores or other implementation/quality problems. In these situations, on-site fidelity assessment holds promise to play a critical role as not only an assessment tool yielding data on fidelity markers, but also as crucial scaffolding for program development and technical assistance. Moreover, occasional in-person contact may bolster trust and confidence with the assessors, which may carry over into remote assessments. Future studies could test the use of remote fidelity assessments within this context by testing for differential impact on alternative measures of program quality or outcome.

Limitations

Because assessor time in our study involved some doctoral level staff and advanced master's level assessors, perhaps at higher personnel costs compared to some statewide implementation staffing, the cost savings may be less dramatic when applied to other settings. However, these concerns are mitigated by our subsequent analyses using our least expensive assessor. We encourage program administrators to apply their own staff base salaries and fringe rates to estimate costs. Moreover, participating VA staff in our study may be partially shielded from the costs of on-site assessment—VA is not a fee-for-service environment in which loss of billable time would affect the agency's bottom-line or staff productivity requirements to the same extent as seen in non-VA providers. Also, participating VAMCs did not pay the costs of the on-site assessments. As detailed in the results section, our travel costs were high (e.g., most on-site assessments required air travel) due to sampling from VA sites nationwide. These

travel costs might not be applicable for statewide use of on-site fidelity assessments. However, results remained significant when excluding travel costs.

The generalizeability for remote fidelity costs methods using the DACTS measure may be limited because it contains many items rated on objective program structures and functions more amenable to assessment via tables of information. Other ACT fidelity assessments requiring extensive observation and interview data to rate clinically-oriented items may be more difficult to adapt for remote assessment (Monroe-DeVita, Teague, & Moser, 2011). However, fidelity assessment methods of clinical competence for other practices have been developed that are amenable to audio-recordings and subsequent assessment (Haddock et al., 2001; McGuire et al., 2012; McGuire et al., in press) or monitoring fidelity via natural language processing of administrative data (Shiner et al., 2013).

Finally, for our qualitative feedback, VA case management team leaders had never had on-site assessments and, therefore, might have a more favorable impression of these methods due to their novelty. These preferences may not generalize to other samples of team leaders who routinely experience on-site visits of various kinds. The costs for on-site reviews of experienced teams may decrease over time.

Table 1. Costs of fidelity assessment methods

	On-site (O) N=19		Phone (P) N=13		Expert Scored Self-report (E) N=32		Statistical tests on cost differences for main categories
	Minutes Mean (SD) Range	Cost in USD Mean (SD) Range	Minutes Mean (SD) Range	Cost in USD Mean (SD) Range	Minutes Mean (SD) Range	Cost in USD Mean (SD) Range	t, df, p value**
Chart review prep by teams	153 (102) (50, 360)	114 (72) (27, 270)	152 (101) (70, 420)	131 (72) (70, 263)	153 (100) (50, 420)	120 (71) (27, 270)	--
Other table prep by teams	611 (586) (100, 2430)	485 (468) (59, 1594)	249 (227) (0, 745)	215 (217) (0, 742)	474 (509) (0, 2430)	382 (409) (0, 1594)	--
All other team assessment time	232 (93) (120, 480)	194 (83) (89, 359)	118 (57) (68, 240)	100 (37) (68, 169)	---	---	--
Total team burden	**984 (645) (340, 3015)**	**782 (510) (225, 1983)**	**510 (248) (250, 1020)**	**439 (200) (251, 887)**	**606 (532) (180, 2760)**	**488 (418) (105, 1813)**	O vs P:t(25)=2.7, p=0.01 O vs E:t(18)=10.2, p<.01 P vs E:t(12)=9.8, p<.01
Total assessor burden*	1268 (295) (465, 1685)	881 (283) (284, 1509)	163 (49) (118, 305)	132 (40) (101, 248)	63 (25) (28, 128)	65 (27) (28, 134)	O vs P:t(19)=11.4, p<.01 O vs E:t(18)=12.7, p<.01 P vs E:t(14)=4.8, p<.01
TOTAL personnel burden (team + assessor)	**2253 (712) (965, 4245)**	**1663 (588) (686, 3083)**	**674 (256) (368, 1170)**	**571 (204) (364, 998)**	**669 (529) (208, 2800)**	**553 (417) (183, 1856)**	O vs P:t(24)=7.5, p<.01 O vs E:t(18)=14.2, p<.01 P vs E:t(12)=8.8, p<.01
Travel costs	**---**	916 (370) (90, 1609)	**---**	**---**	**---**	**---**	**--**
GRAND TOTAL COSTS	**---**	**2579 (804) (776, 4278)**	**---**	**571 (204) (364, 998)**	**---**	**553 (417) (183, 1856)**	O vs P:t(21)=10.4, p<.01 O vs E:t(18)=13.7, p<.01 P vs E: t(12)=8.8, p<.01

*Total assessor burden was computed by averaging the time and costs of two assessors for all phone and self-report methods as well as some on-site visits that included two assessors to ensure reliability.

**T-values and degrees of freedom (df) were adjusted using the Satterwaithe correction when variances between groups were not equivalent. Independent t-tests were used when comparing phone and on-site costs (distinct sites used for each fidelity method), while paired t-tests were used for analyses involving export-scored self-report (repeated measures within sites).

SD= standard deviation

References

Aarons, G. A., Sommerfeld, D. H., Hecht, D. B., Silovsky, J. F., & Chaffin, M. J. (2009). The impact of evidence-based practice implementation and fidelity monitoring on staff turnover: evidence for a protective effect. *J Consult Clin Psychol, 77*(2), 270-280. doi: 10.1037/a0013223

Bond, G. R., Drake, R. E., McHugo, G. J., Peterson, A. E., Jones, A. M., & Williams, J. (2014). Long-term sustainability of evidence-based practices in community mental health agencies. *Adm Policy Ment Health, 41*(2), 228-236. doi: 10.1007/s10488-012-0461-5

Bond, G. R., Drake, R. E., Mueser, K. T., & Latimer, E. (2001). Assertive community treatment for people with severe mental illness: Critical ingredients and impact on patients. *Disease Management & Health Outcomes, 9*, 141-159.

Bond, G. R., McGrew, J. H., & Fekete, D. M. (1995). Assertive outreach for frequent users of psychiatric hospitals: a meta-analysis. *Journal of Mental Health Administration, 22*(1), 4-16.

Coldwell, C. M., & Bender, W. S. (2007). The effectiveness of assertive community treatment for homeless populations with severe mental illness: a meta-analysis. *American Journal of Psychiatry, 164*(3), 393-399.

Crabtree, B. F., & Miller, W. L. E. (1999). *Doing qualitative research: Second edition*. Thousand Oaks: Sage Publications.

Cuddeback, G. S., Morrissey, J. P., Domino, M. E., Monroe-DeVita, M., Teague, G. B., & Moser, L. L. (2013). Fidelity to recovery-oriented ACT practices and consumer outcomes. *Psychiatric Services, 64*(4), 318-323. doi: 10.1176/appi.ps.201200097

Drake, R. E., Bond, G. R., & Essock, S. M. (2009). Implementing evidence-based practices for people with schizophrenia. *Schizophrenia Bulletin, 35*(4), 704-713. doi: 10.1093/schbul/sbp041

Haddock, G., Devane, S., Bradshaw, T., McGovern, J., Tarrier, N., Kinderman, P., . . . Harris, N. (2001). An investigation into the psychometric properties of the cognitive therapy scale for psychosis (CTS-Psy). *Behavioural and Cognitive Psychotherapy, 29*, 221-233.

Herdelin, A. C., & Scott, D. L. (1999). Experimental Studies of the Program of Assertive Community Treatment (PACT): A Meta-Analysis. *Journal of Disability Policy Studies, 10*(1), 53-89. doi: 10.1177/104420739901000105

Latimer, E. (1999). Economic impacts of assertive community treatment: A review of the literature. *Canadian Journal of Psychiatry, 44*, 443-454.

McGrew, J. H., Bond, G. R., Dietzen, L. L., & Salyers, M. P. (1994). Measuring the fidelity of implementation of a mental health program model. *Journal of Consulting and Clinical Psychology, 62*(4), 670-678.

McGrew, J. H., Stull, L. G., Rollins, A. L., Salyers, M. P., & Hicks, L. J. (2011). A comparison of phone-based and on-site assessment of fidelity for assertive community treatment in Indiana. *Psychiatr Serv, 62*(6), 670-674. doi: 10.1176/appi.ps.62.6.670

McGrew, J. H., White, L. M., Stull, L. G., & Wright-Berryman, J. (2013). A comparison of self-reported and phone-administered methods of ACT fidelity assessment: a pilot study in Indiana. *Psychiatr Serv, 64*(3), 272-276. doi: 10.1176/appi.ps.001252012

McGuire, A. B., Stull, L. G., Mueser, K. T., Santos, M., Mook, A., Rose, N., . . . Salyers, M. P. (2012). Development and reliability of a measure of clinician competence in providing illness management and recovery *Psychiatric Services, 63*(8), 772-776.

McGuire, A. B., White, D. A., Bartholomew, T., Flanagan, M. E., McGrew, J. H., Rollins, A. L., . . . Salyers, M. P. (in press). The Relationship Between Provider Competence, Content Exposure, and Consumer Outcomes in Illness Management and Recovery Programs. *Administration and Policy in Mental Health and Mental Health Services Research*, 1-11.

McHugo, G. J., Drake, R. E., Teague, G. B., & Xie, H. (1999). Fidelity to assertive community treatment and client outcomes in the New Hampshire Dual Disorders Study. *Psychiatric Services, 50*(6), 818-824.

McHugo, G. J., Drake, R. E., Whitley, R., Bond, G. R., Campbell, K., Rapp, C. A., . . . Finnerty, M. T. (2007). Fidelity outcomes in the national implementing evidence-based practices project. *Psychiatric Services, 58*(10), 1279-1284.

Monroe-DeVita, M., Teague, G. B., & Moser, L. L. (2011). The TMACT: A new tool for measuring fidelity to assertive community treatment. *Journal of the American Psychiatric Nurses Association, 17*(1), 17-29.

Mueser, K. T., Bond, G. R., Drake, R. E., & Resnick, S. G. (1998). Models of community care for severe mental illness: A review of research on case management. *Schizophrenia Bulletin, 24*, 37-74.

Phillips, S. D., Burns, B. J., Edgar, E. R., Mueser, K. T., Linkins, K. W., Rosenheck, R. A., . . . McDonel Herr, E. C. (2001). Moving assertive community treatment into standard practice. *Psychiatric Services, 52*(6), 771-779.

Rollins, A. L., McGrew, J. H., Kukla, M., McGuire, A. B., Flanagan, M. E., Hunt, M. G., . . . Salyers, M. P. (2016). Comparison of Assertive Community Treatment Fidelity Assessment Methods: Reliability and Validity. *Adm Policy Ment Health, 43*(2), 157-167. doi: 10.1007/s10488-015-0641-1

Shiner, B., D'Avolio, L. W., Nguyen, T. M., Zayed, M. H., Young-Xu, Y., Desai, R. A., . . . Watts, B. V. (2013). Measuring use of evidence based psychotherapy for posttraumatic stress disorder. *Adm Policy Ment Health, 40*(4), 311-318. doi: 10.1007/s10488-012-0421-0

Teague, G. B., Bond, G. R., & Drake, R. E. (1998). Program fidelity in assertive community treatment: development and use of a measure. *American Journal of Orthopsychiatry, 68*(2), 216-232.

van Vugt, M. D., Kroon, H., Delespaul, P. A. E. G., Dreef, F. G., Nugter, A., Roosenschoon, B.-J., . . . Mulder, C. L. (2011). Assertive community treatment in the Netherlands: outcome and model fidelity. *Canadian Journal of Psychiatry, 56*(3), 154-160.

Ziguras, S., & Stuart, G. (2000). A meta-analysis of the effectiveness of mental health case management over 20 years. *Psychiatric Services, 51*, 1410-1421.

Supplemental Table: Fidelity assessment method cost comparison when applying the study's lowest Master's level assessor salary and fringe to cost calculations

	On-site Cost in USD N=19 Mean (SD) Range	Phone Cost in USD N=13 Mean (SD) Range	Self-report Cost in USD N=32 Mean (SD) Range	Statistical tests on cost differences t, df, p value t, df, p value**
Total assessor burden*	775 (180) (284, 1029)	100 (30) (72, 186)	38 (15) (17, 78)	O vs P:t(19)=16.0, p<.01 O vs E:t(18)=17.9, p<.01 P vs E: t(12)=6.8, p<.01
TOTAL personnel burden (team + assessor)	**1557 (570) (686, 2937)**	**539 (204) (334, 969)**	**526 (416) (159, 1838)**	O vs P:t(24)=7.1, p<.01 O vs E: t(18)=17.4, p<.01 P vs E:t(12)=10.0, p<.01
Travel costs	916 (370) (90, 1609)	**---**	**---**	**--**
GRAND TOTAL COSTS	**2472 (769) (776, 4132)**	**539 (204) (334, 969)**	**526 (416) (159, 1838)**	O vs P:t(22)=10.4, p<.01 O vs E: t(18)=14.6, p<.01 P vs E: t(12)=10.0, p<.01

*Total assessor costs were computed by averaging the time of two assessors for all phone and self-report methods as well as some on-site visits that included two assessors to ensure reliability, and applying salary and fringe costs of our lowest salaried, Master's level assessor.
**T-values and degrees of freedom (df) were adjusted using the Satterwaithe correction when variances between groups were not equivalent. Independent t-tests were used when comparing phone and on-site costs (distinct sites used for each fidelity method), while paired t-tests were used for analyses involving export-scored self-report (repeated measures within sites).

SD= standard deviation
O=on-site
P=phone administered
E=expert-scored self-report